Abstract:
Millions of people live in this world with incapacities of understanding the environment due to visual impairment. According to the World Health Organization, there are approximately 285 million people who have visual impairments, 39 million of them are blind and 246 million have a decrease of visual acuity. Although they can develop alternative approaches to deal with daily routines, they also suffer from navigation difficulties as well as social awkwardness. It is very difficult for them to adjust to an unfamiliar environment. Computer vision technologies, particularly the Deep Convolutional Neural Network, were developed rapidly in recent years. Use of the state-of-the-art computer is promising Vision techniques to help vision loss sufferers. In this project we used the sense of hearing in order to visualize the object kept before the person and the camera. We used the state of the art “You Only Look Once: Unified, Real-Time Object Detection” YOLOv4 algorithm trained on the COCO dataset to identify the object present before the person. We calculate the spatial location of the object in the frame and have the output of the result as an audio speech so that the visually impaired person could hear.

Keywords: Object detection, Yolov4, voice feedback.

I. INTRODUCTION

In recent years computer vision technologies have been developed which are very accurate and give promising results presented real-time object detection system using a CNN in order to recognize objects. In this project, we explored the possibility of using the hearing sense to understand visual objects. The sense of sight and hearing sense share a striking similarity. We built a real-time object detection with a voice feedback system with the goal of telling the user what all is in the surroundings with its spatial position. We developed a system to be used in an Android-based mobile phone by the visually impaired for detecting objects in their vicinity, which will help them move around safely, without crashing into any object. The detection of objects will be done from real-time video taken from their mobile phone camera. We used the state of the art “You Only Look Once: Unified, Real-Time Object Detection” YOLOv4 algorithm trained on the COCO dataset to identify the object present before the person. Then the label of the object and its location in the frame is identified and then converted into audio by using Text to Speech conversion which will be the anticipated output. The person can use the output of our system in order to make him aware of his surroundings. In this project, we used the sense of hearing in order to visualise the object kept before the person and the camera.

II. LITERATURE REVIEW

The first Deep Learning object detector model was called the Over feat Network which used Convolutional Neural Networks (CNNs) along with a sliding window approach. It classified each part of the image as an object/nonobject and subsequently combined the results Object detector models have gone through various changes throughout the years since 2012. The first breakthrough in object detection was the RCNN which resulted in an improvement of nearly 30% over the previous state-of-the-art. The RCNN Model was a highly influential model that has shaped the structure of modern object detectors. It was the first detector that proposed the two-step approach [1]. The Fast RCNN came out soon after the RCNN and was a substantial improvement upon the original.

The Fast RCNN is also a two-step model which is quite similar to the RCNN, in that it uses selective search to find some regions and then runs each region through the object detector network. The Faster RCNN came out soon after the Fast RCNN paper. It was meant to represent the final stage of what the RCNN set out to do. It proposed a detector that was learned end to end. Selective Search was serviceable but took a lot of time and set a bottleneck for accuracy. Single Shot MultiBox Detector came out in 2015, boasting state-of-the-art results at the time and real-time speeds. The SSD uses anchors to define the number of default regions in an image. The major difference between the SSD from other architectures is that it was the first model to propose training on a feature pyramid.

The YOLO group of architectures were constructed in the same vein as the SSD architectures. The image was run through a few convolutional layers to construct a feature map. The concept of anchors was used here too, with every grid cell acting as a pixel point on the original image. The YOLO algorithm-generated 2 anchors for each grid cell. Unlike the Fast RCNN, Yolo has only one head [2].

YOLOv4 is a real-time object detection model published in April 2020 that achieved state-of-the-art performance on the COCO dataset. It works by breaking the object detection task into two pieces, regression to identify object positioning via
bounding boxes and classification to determine the object's class. This implementation of Yolov4 uses the Darknet framework [3].

By using YOLOv4, you are implementing many of the past research contributions in the YOLO family along with a series of new contributions unique to YOLOv4 including new features: WRC, CSP, CmBN, SAT, Mish activation, Mosaic data augmentation, CmBN, DropBlock regularization, and Clou loss. YOLOv4 uses a better object detection network architecture and new data augmentation techniques. YOLOv4 has an incredibly high performance for a very high FPS, this is a major improvement from previous object detection models which only had either high performance or high inference speeds [3].

Figure.1. Yolov4 architecture

III. EXISTING WORK

There are various existing models that does real time object detection with voice feedback but the major drawback with these models is that they use older algorithm for object detection like Efficient Det, R-CNN, ATSS, ASFF or YOLOv3. The problem with these algorithms is that the accuracy of the object detection and its real time speed does not go hand in hand, if the accuracy is good then the real time speed is slow and vice versa.

IV. PROPOSED WORK

In the proposed system, we are using YOLOv4 as the main difference from the existing system. YOLOv4 is twice as fast as EfficientDet with comparable performance. In addition, compared with YOLOv3, the AP and FPS have increased by 10% and 12%, respectively.

Other than that, we are using Python3 for this project, the camera is initialized by using the OpenCV library and the camera starts capturing frames with the rate of 30 frames per second and feeds them to the algorithm. Then the system uses YOLOv4 which is trained on the COCO dataset and Dark Neural Network (DNN) to identify the object kept before the user. The object identified is later converted to an audio segment using text to speech conversion.

The audio segment is the output of our system that gives the spatial location and name of the object to the person. Now by using this information the person can have a visualization of the objects around him. The proposed system will even protect the person from colliding with the objects around him hence securing him from injuries.

V. SYSTEM DESIGN

The entire system is present as an Android based smartphone application.

To start with, the visually impaired user starts by opening the camera and the video is fed in real time. Each frame goes under the YOLOv4 algorithm and object and location detection is carried out. The description of the detected output is generated. The backend system creates a thread in background and captures the description of the detected output. In Parallel, the other frames in the video are running in the foreground. This result is fed to a text to speech conversion and finally the user gets the audio feedback. Once the thread finishes the feedback, it captures the latest description from the detected output. This process is asynchronous and tends to be fast. Hence there tends to be no lag.

V. SYSTEM IMPLEMENTATION

1. ANDROID APPLICATION

When the user decides to start the object detection process, he opens the Camera View of the Android App. This triggers the object detection process. The video frames are pre-processed to ensure that the frames do not have noise disturbance, the frames are not blurry, etc. The individual video frames are sent asynchronously to the YOLO object detection system.

2. OBJECT DETECTION

To use YOLO algorithm, it is necessary to establish what is actually being predicted. Ultimately, we aim to predict a class of an object and the bounding box specifying object location. Each bounding box can be described using four descriptors:

1. centre of a bounding box \((bx, by)\)
2. width \((bw)\)
3. height \((bh)\)
4. value \(c\) is corresponding to a class of an object (such as: car, traffic lights, etc.).
In addition, to predict the \( p_c \) value, which is the probability that there is an object in the bounding box.

\[
y = (p_c, h_x, h_y, h_w, h_a, c)
\]

\( b_y \\
( b_i, b_j ) \)

**Figure.3. Object detection**

The image is split into cells, typically using a 19×19 grid. Each cell is responsible for predicting 5 bounding boxes (in case there is more than one object in this cell). Most of these cells and bounding boxes will not contain an object. Therefore, we predict the value \( p_c \), which serves to remove boxes with low object probability and bounding boxes with the highest shared area in a process called non-max suppression.

**Figure.4. Non-max suppression**

3. AUDIO OUTPUT

The audio system converts the object’s label and location into audio format. The audio is then played on the smartphone speaker as an output for the user.

VI. CONCLUSION AND FUTURE WORK

This project intended to achieve real-time Object Detection using Yolov4 with Voice Feedback. Our project will help the visually impaired as we were able to detect objects more accurately and independently identify objects with the exact location of an object in the x, y-axis image which is conveyed in audio format using text-to-speech. Using small Deep Neural Network architectures for object detection such as Tiny YOLO are giving good results and show that they can be used for real-time object detection using mobile devices. Thus, the technologies of computer vision and deep learning can be successfully used to develop a real-time app for the visually impaired. It will enable them to detect objects in their vicinity thus helping them in better navigation. We can further enhance this model by adding a Facial Recognition system to it, which will help identify people in an image and relay it to a visually challenged person using the app. This model can also be trained using PP-YOLO which is another new YOLO upgrade based on a deep learning framework called PaddlePaddle, and it improves the YOLO v4 model to obtain a better balance between effectiveness and efficiency. PP-YOLO can achieve an inference speed of 72.9 FPS, which is also higher than 65 FPS of YOLOv4 and would thus make out the existing model faster.

VII. REFERENCES

